DOCUMENT RESUME

ED 202 911 TM 810 391

AUTHOR Weber, Larry J.; Frary, Robert B.
TITLE Profile Uniqueness in Student Ratings of Instruction.
PUB DATE 16 Apr 81
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, Ca, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cluster Analysis; Evaluation Methods; Higher Education; Statistical Analysis; *Student Evaluation of Teacher Performance; Test Interpretation
IDENTIFIERS Likert Scales

ABSTRACT

An approach to partitioning the variance in student ratings not previously reported in the literature is described. The new approach provides an alternative basis for interpreting faculty evaluations that overcome objections to current practices. Student responses on evaluation forms were cluster-analyzed to establish homogeneous subgroups of students within classes who exhibited distinct evaluation profiles. All classes were subjected to profile analysis for parallels across groups. Results caused investigators to seriously question the continued use of traditional methods of interpreting student ratings on the basis of averaging responses across each item and reporting the resulting means. The investigators recommend reporting profiles from clusters of student ratings as far superior to the typical reporting of evaluation results. That is, it is more sensible to emphasize the existence of differing rating constituencies than to obscure them through averaging. An approach such as the one proposed would prevent invidious comparisons between instructors based on small scale differences and would provide both administrators and instructors with much more concrete bases for decision making. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED202911

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

PROFILE UNIQUENESS IN STUDENT RATINGS OF INSTRUCTION

Larry J. Weber

and

Robert B. Frary

Virginia Polytechnic Institute and State University

Blacksburg, Virginia

April 16, 1981

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. J. Weber

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, California, April 13-17, 1981.

TM 810 391

PROFILE UNIQUENESS IN STUDENT RATINGS OF INSTRUCTION

Larry J. Weber and Robert B. Frary
Virginia Polytechnic Institute and State University

Student ratings of instruction are widely used in colleges and universities to provide information for improvement of instruction and for personnel decisions, such as those related to tenure, promotion and salary adjustment. The appropriateness of this usage is frequently questioned on the basis that the resulting scores applied to faculty members lack criterion-related validity or that a substantial proportion of their variance is attributable to student or course variables. Empirical investigations have yielded contradictory results but, for the most part, have provided little evidence that current usage of ratings is highly inappropriate.

The present paper describes an approach to partitioning the variance in student ratings not reported previously. Moreover, empirical results from use of this approach provide evidence that seriously questions the continuing use of traditional methods of interpreting student ratings. At the same time the new approach provides an alternative basis for interpretation that overcomes many objections to current practices.

Description of New Approach

It is not difficult to imagine the existence of groups within a class who react differently with respect to a questionnaire item. For example, consider the rating of effectiveness of communication of an instructor with a strong regional accent. Students from outside the region might base their

ratings almost entirely on the extent to which the accent made understanding difficult, while students sharing the same accent would base their ratings on other aspects of communication.

Suppose that the "outsiders" give generally low ratings for communication and that the others give high ratings. The usual averaging process would yield a mediocre mean rating, which at best would be uninformative and at worst untrue or misleading. Seeking to improve this rating, the instructor might or might not discover the problem, depending on whose advice was sought.

Analysis of responses across several items can be done in a manner which displays rather than obscures situations such as that just described. This goal can be met simply by clustering respondents who give the same or nearly the same ratings across all items of a questionnaire. For each cluster so established, a profile may be drawn across items based on the mean intercluster response to each item. Figure 1 shows hypothetical mean ratings from three homogeneously responding clusters on a six-item questionnaire with a four-point scale. Clusters 1 and 2 view the instructor as performing more or less uniformly well across the six items, but Cluster 2 has a distinctly lower opinion of this performance than Cluster 1. Clusters such as these will be referred to as parallel. They may well be rating the instructor from the same viewpoint but with differing standards of excellence.

If the clusters constituting a class were essentially parallel, averaging responses across all students would not be inappropriate. However, existence of clusters such as Cluster 3 in Figure 1 argue strongly against this procedure. While Clusters 1 and 2 rate the instructor uniformly across all items, Cluster 3 gives relatively much lower ratings on items 2, 5 and 6. This departure from parallelism suggests the need for further investigation

in these areas, which might well have been obscured under the usual averaging procedures, especially if Cluster 3 were relatively small.

There are various problems and limitations in implementing the foregoing ideas. A satisfactory method for clustering according to response similarity must be adopted and standards for parallelism must be established. The questionnaire must be rather brief to foster the existence of clusters of reasonable size with relatively homogeneous responses across all items. Further, implementation could lead to the conclusion that in most classes, the response clusters are parallel, which would suggest that the hypothesized benefits of the new approach are largely nonexistent.

In what follows, one approach to implementation is presented along with results suggesting the extent to which nonparallel groups may exist.

Implementation

Twenty undergraduate classes with a minimum size of 30 students were randomly selected from all undergraduate sections of a typical college within a comprehensive university in which the evaluation questionnaire, Student Perceptions of Instruction (Table 1), had been administered in the fall of 1979. The questionnaire elicited responses on a four-point Likert scale to six items pertaining to the quality of instruction. Only students who responded to all six items were included in the analyses which follow. This restriction was necessary to permit the required analyses. Fewer than 10% of the students were dropped from any class.

To establish profile groups, the responses of students in each class were cluster-analyzed using Johnson's MAX procedure (Baker, 1972). The criterion adopted was that a maximum intercluster distance across all six items on the evaluation questionnaire could not exceed three scale points. In other words,

students in a cluster were within a half point of each other on the average across all six questions. For instance, if one student rated the instructor excellent (4 points) on all six items of the scale and another student rated him excellent on three items and good (3 points) on three items, these students could be members of the same cluster. A criterion was established such that at least five students had to find membership in the same cluster to be acknowledged as a group of meaningful size. Then for each such cluster, a profile was determined by averaging responses of the cluster for each item.

Parallelness of two clusters was evaluated by summing the absolute values of slope differences between successive questionnaire items. To illustrate this concept, parallelness of Groups 1 and 2 of Figure 1 will be evaluated. From item 1 to item 2, or from item 2 to item 3 there is no slope difference. However, from item 3 to item 4, Cluster 1 gains .2 and Cluster 2 gains .7 for a difference of .5. Between item 4 and item 5, Cluster 1 loses .2 and Cluster 2 loses .5 for a difference of .3. Between item 5 and item 6, Cluster 1 loses .2 and Cluster 2 shows no change for a difference of .2. The total difference across the five between-item slopes is 1.0. Arbitrarily a maximum of 1.25 was established for defining parallel profiles. This criterion would, for example, permit two groups to vary in any manner within a band of width .25.

Under this standard, Cluster 3 is not parallel to either of the others. The sum of the slope difference between the profiles of Clusters 1 and 3 is 4.1, and between Clusters 2 and 3 it is also 4.1. The situation depicted in this example suggests that Cluster 3 is rating the instructor on some basis other than that of Cluster 1 and 2, in which case it might be hazardous to draw conclusions about instruction without taking the profile differences into account.

Hypothetical Cluster 1 gave the instructor high ratings on all six items. When the data from the 20 classes were analyzed, the occurrence of such a phenomenon was sufficiently frequent to justify special attention. Profiles similar to the Cluster 1 profiles in Figure 1 will be referred to as uniform praise clusters. In order for a cluster to be assigned to the uniform praise category two criteria were established. Specifically, no mean rating below 3.3 may have been earned on any of the six scales, and the average across scales must have been 3.75 or above.

Given the arbitrary nature of the definitions and criteria just established, some sort of inferential statistical test seems called for to determine the extent to which observed outcomes might represent chance occurrences. The situation here is somewhat tenuous. The populations represented by the samples are of uncertain definition, and the samples are not random insofar as individual students are concerned. Nevertheless, populations can be envisioned as consisting of all students who will take similar courses from the instructor in question, and there is no reason to believe any of the samples was biased. Having accepted this basis for sampling, many statisticians would then think naturally of a test for parallelness of profiles such as that described by Morrison (1967). Unfortunately, any parallelness test based on within-group slope variation is quite inapplicable. This statement follows because the basis for group formation suppresses within-group variation not only of responses to an item but also of the item-to-item differences which define the slopes analyzed to evaluate parallelism inferentially.

Actually, the real inferential issue is that of whether the clusters, however derived, have additional potential members in the population to the extent that new samples should be expected to contain reasonable numbers of them. Investigations into this phenomenon have been somewhat sparse.

However, a fairly comprehensive study by Edelbrock (1979) suggests that only large clusters with small maximum intercluster distances are likely to represent population subgroups which consistently yield similar clusters in repeated samples. As in the present study, Edelbrock did not enter all subjects into clusters when doing so would violate a preestablished maximum intercluster distance.

Results

Table 2 summarizes the major findings of the study. Inspection of these outcomes suggests the following observations.

The number of students who attained membership in a cluster varied considerably. With a criterion of a minimum of five students having to rate an instructor similarly in order for a cluster to exist, the percentage of students who were able to find membership in a cluster ranged from 0% (Class 20) to 94% (Class 13). The median percentage of students who could be grouped in a cluster was 73%, which suggests that approximately 3/4 of the students in a class will rate the instructor in a fashion similar to others in a group of at least five students. (Refer to Column 3 in Table 2).

The number of clusters (Column 4) appears to be related to the size of the class, with the exception of Class 1, which has a very high average cluster size. It was noteworthy that Class 20 had no clusters at all. For the other 19 classes, the number of clusters ranged from two to eight. Such information suggests that, in a class consisting of thirty students, one might expect to find from two to five distinct groups whose rating profiles differ from one another. Moreover, the information contained in Column 6 suggests that a dominant single group is virtually never present. While the percentage of students in the largest single cluster varied from 18% to 61%, the median percentage of students in the largest group is 26.5%.

Information contained in Column 7 pertains to the presence of a uniform praise group within a class. Twelve of the twenty classes (60%) contained clusters of students which met the two criteria for uniform praise. The variability of the size of the uniform praise groups was considerable, ranging from 10% (Class 9) to 61% (Class 12). The median was 27.5%.

Columns 8 and 9 present data reflecting the major concern of this study, the profile parallelism of student ratings of classroom instruction. Only ten of the twenty classes contained parallel clusters, with four of them having two sets of parallel profiles. In only two instances were there three clusters that were parallel to one another.

Inspection of the data can only lead to the conclusion that response patterns of individual students and clusters are highly idiosyncratic. Of the 74 distinct clusters for all twenty classes only 30 (41%) were parallel to one or two other clusters. These 30 clusters comprised 304 of the total number of 1015 students in the study, only 30% of the sample. Finally, while the average number of clusters in a class was almost four, the average number of parallel clusters was less than two.

Conclusions and Recommendations

The results of this study suggest that it may be generally inappropriate to use student evaluations either for inproving instruction or for personnel decisions, if they are interpreted on the basis of averaging responses across each item and reporting the resulting means. To the extent that clusters with nonparallel profiles exist, it would be wrong to advise an instructor to modify his teaching on the basis of some mean value. Such a change might alienate one group of students while pleasing another. Moreover,

what is known about the extent to which clusters in a sample are likely to represent existing population groups (Edelbrock, 1979) suggests that substantial caution be observed in attempting to improve perceptions of some presumed constituency. Only if a cluster is numerically large or contains a substantial proportion of a class might there be immediate justification for adapting instruction to that group's desires or needs. In the 20 classes sampled for this study, only 12 contained one or more clusters constituting as much as 25% of the class each, and only four contained clusters with 20 or more students per cluster.

To give rewards to faculty members on the basis of average ratings over an entire class would seem equally as inappropriate as using average ratings to make instructional decisions. A set of mediocre mean ratings may obscure any number of substantial strengths and weaknesses which administrators might well choose to reward or penalize if their existence were known.

In view of the foregoing results and conclusions, the investigators recommend reporting profiles from clusters from student ratings as far superior to the typical reporting of evaluation results. It makes much more sense to emphasize the existence of differing rating constituencies than to obscure them through averaging.

Of course, administrative use of the profiles would be more taxing than present practices, which typically rank instructors on a single value, the grand average across all questions. Each set of profiles would have to be inspected for evidence of especially good or bad performance. In the manner of a Q-sort an administrator or a panel could classify the profiles from a class as good, mediocre or bad. For example, the presence of a large proportion of the class in a uniform praise group would be evidence of good teaching. The presence of even a rather small group giving uniformly

bad ratings might be evidence of unsatisfactory teaching (alienation of students). Another factor to be considered is the proportion of students who do not attain cluster membership. A high proportion might indicate erratic instructional performance or some other unusual circumstance.

Such an approach would prevent invidious comparisons between instructors based on small scale differences and would provide both administrators and instructors with much more concrete bases for decision making.

References

Baker, F.B. Numerical taxonomy for educational researchers. Review of Educational Research, 1972, 42, 435-358.

Morrison, D.F. Multivariate Statistical Methods. New York: McGraw Hill Book Co., Inc., 1967

Edelbrock, C. Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. J. Multivariate Behavioral Research, 1979, 14, 367-384.

TABLE 1 -- Questionnaire Content: Student Perceptions of Instruction

Response Scale:
4 = Excellent
3 = Good
2 = Fair
1 = Poor

Items:
1. Apparent knowledge of subject matter.
2. Success in communicating or explaining subject matter.
3. Degree to which subject matter was made stimulating or relevant.
4. Concern and respect for students as individuals.
5. Fairness in assigning grades.
6. Administration of the class and organization of materials.

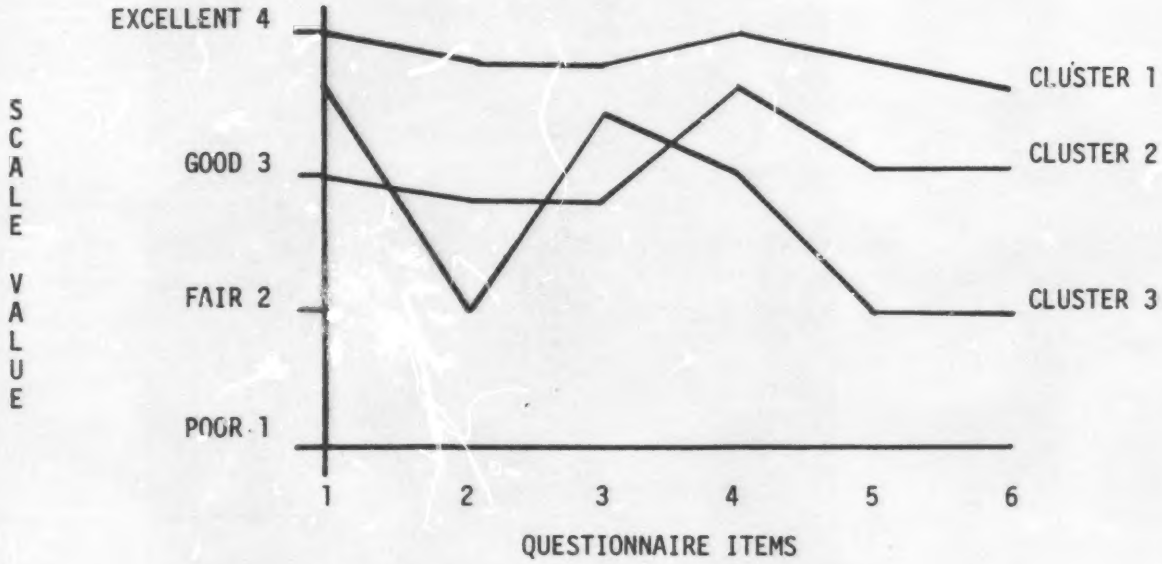

FIGURE 1 PARALLEL AND NON-PARALLEL CLUSTERS

TABLE 2 -- Clustering Statistics for Classes Responding to the Questionnaire, Student Perceptions of Instruction

(1) Class no.	(2) No. students with complete responses	(3) No. students assigned to clusters (prop.)	(4) No. of clusters	(5) Average cluster size	(6) No. in largest cluster (prop. of (2))	(7) No. students in uniform praise cluster (prop. of (2))	(8) No. of parallel clusters	(9) No. students in parallel clusters (prop. of (2))
1	62	54 (.83)	2	27	33 (.53)	33 (.53)	0	-
2	57	40 (.70)	4	8	19 (.33)	0	0	-
3	78	72 (.92)	8	9	21 (.27)	11 (.14)	3	38 (.49)
4	46	31 (.67)	4	8	10 (.22)	0	0	-
5	40	31 (.78)	4	8	16 (.40)	10 (.25)	3	25 (.63)
6	59	43 (.73)	6	7	11 (.19)	8 (.15)	2, 2	16,14 (.27,.24)
7	39	22 (.56)	2	11	11 (.28)	0	0	-
8	84	66 (.79)	7	9	15 (.18)	15 (.18)	2, 2	21,25 (.25,.30)
9	49	28 (.57)	4	7	9 (.18)	5 (.10)	2	14 (.29)
10	42	18 (.43)	3	6	8 (.19)	0	0	-
11	30	27 (.90)	3	9	14 (.47)	14 (.47)	0	-
12	31	26 (.84)	2	13	19 (.61)	19 (.61)	2	26 (.63)
13	32	30 (.94)	3	10	13 (.41)	11 (.34)	2	24 (.75)
14	91	83 (.91)	6	14	27 (.30)	27 (.30)	2, 2	14,37 (.15,.41)
15	72	41 (.57)	3	14	19 (.26)	0	0	-
16	34	26 (.76)	2	13	17 (.50)	17 (.50)	0	-
17	66	46 (.70)	5	9	20 (.26)	6 (.19)	2, 2	11,26 (.17..37)
18	35	18 (.51)	3	6	8 (.23)	0	0	-
19	36	19 (.53)	3	6	8 (.22)	0	2	13 (.36)
20	30	0 (.00)	0	0	0 (.00)	0	0	-